Clearance number PA# 96-701

SEX DIFFERENCES ON U. S. AIR FORCE PILOT SELECTION TESTS[1]

Dr. Thomas R. Carretta
Armstrong Laboratory Human Resources Directorate
Brooks Air Force Base, TX

The study of sex and ethnic group differences is an important issue in the evaluation of personnel measurement instruments. Several principles must be considered when addressing the measurement of group differences including whether the selection instruments measure the same factors for all groups (i.e., factorial invariance), group mean score differences, and differential validity. Federal guidelines prohibit the use of personnel selection tests that show test bias. Evidence of differential prediction is accepted as evidence of test bias.

Several recent U. S. Air Force (USAF) studies of sex differences on pilot selection tests and training performance are reviewed. The issues addressed include invariance of factor structure for selection tests, mean score differences, acquisition of pilot job knowledge and flying skills during training, and prediction of training performance. These studies focus on 2 widely-used USAF pilot selection tests, the Air Force Officer Qualifying Test (AFOQT; Berger, Gupta, Berger, & Skinner, 1990) and the Basic Attributes Test (BAT; Carretta & Ree, 1993).

Measures

Air Force Officer Qualifying Test

The AFOQT is a paper-and-pencil battery used for selection into officer commissioning and aircrew training programs. It has been in use since 1957 and new forms are developed about every 7 years. The current form has 16 tests that are combined into 5 composites (Verbal, Quantitative, Academic Aptitude [Verbal + Quantitative], Pilot, and Navigator-Technical; Berger et al., 1990). The AFOQT Pilot composite is a component of a USAF pilot selection composite implemented in June 1993 known as the Pilot Candidate Selection Method (PCSM; Carretta, 1992). The 16 tests are Verbal Analogies (VA), Arithmetic Reasoning (AR), Reading Comprehension (RC), Data Interpretation (DI), Word Knowledge (WK), Math Knowledge (MK), Mechanical Comprehension (MC), Electrical Maze (EM), Scale Reading (SR), Instrument Comprehension (IC), Block Counting (BC), Table Reading (TR), Aviation Information (AI), Rotated Blocks (RB), General Science (GS), and Hidden Figures (HF).

Basic Attributes Test

The BAT is a computer-based battery used for pilot selection (Carretta & Ree, 1993). It has 5 tests that measure psychomotor coordination, short-term memory, and attitude toward risk-taking. It was operationally implemented in June 1993 and contributes to the PCSM composite (Carretta, 1992). The BAT tests are Two-Hand Coordination, Complex Coordination, Item Recognition, Time Sharing, and Activities Interest Inventory.

STUDIES OF SEX DIFFERENCES ON USAF PILOT SELECTION TESTS

Factor Structure

Confirmatory factor analytic models were compared for men and women to determine the similarity of their respective factor structures. Factorial invariance is demonstrated when the factor loadings are the same for the groups being compared (McArdle, 1996). A χ^2 test was done to determine if the loadings for a score on a factor were the same for both sexes (Bentler, 1989).

AFOQT. Carretta and Ree (1995) examined the AFOQT factor structure for 219,887 male and 50,081 female USAF officer applicants. The factor structure tested for men and women had been confirmed previously by Carretta and Ree (1996; see Figure 1). The hierarchical factor is general cognitive ability (*g)* and the 5 lower-order factors are verbal, math, spatial, aviation interest/aptitude, and perceptual speed.

---

[1] Previously published as Carretta, T. R. (1997). Sex differences on U. S. air force pilot selection tests. *Proceedings of the Ninth International Symposium on Aviation Psychology*, Columbus, OH, 1292-1297.

In spite of group mean score differences on the 16 tests, the Carretta and Ree (1996) model showed good fit for both sexes and the proportions of total and common variance accounted for by each factor were similar. These results were interpreted as evidence of near identity of cognitive structure for men and women and were consistent with a similar study of the Armed Services Vocational Aptitude Battery (ASVAB; Ree & Carretta, 1995).

BAT. In a sample of 354 USAF enlisted personnel, Ree and Carretta (1994) examined the factor structure of the BAT psychomotor tests in the presence of 4 *g*-loaded verbal and math tests. A confirmatory factor analysis yielded general cognitive and general psychomotor factors, 3 lower-order psychomotor factors, and 2 lower-order cognitive factors. Contrary to expectations, the psychomotor tests contributed to the general cognitive factor. Based on these results, the factor structure of the operational BAT cognitive, psychomotor, and attitude toward risk scores was examined for men and women in the presence of the AFOQT verbal and math tests. The AFOQT tests were included in the analyses to examine the relations between the BAT scores and measures of *g*. The samples consisted of 4,888 male and 465 female USAF pilot applicants tested on the BAT and AFOQT.

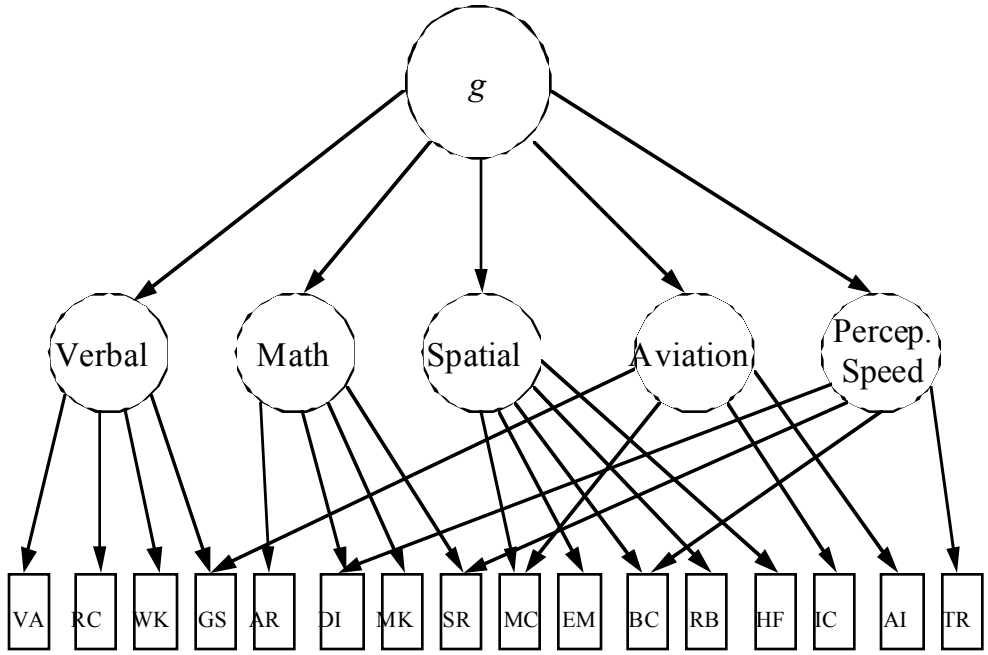

Figure 1. Air Force Officer Qualifying Test Factor Structure

The model (Figure 2) was based on results from earlier confirmatory factor analyses of the AFOQT (Carretta & Ree, 1996) and the BAT psychomotor tests (Ree & Carretta, 1994). The model included general cognitive and general psychomotor factors and the lower-order factors of verbal, math, two-hand coordination, complex coordination, response time, time sharing, and activities interests. All scores contributed to *g*. The model showed good fit and the proportion of common and total variance accounted for by the factors was similar for both sexes.

Mean Score Performance

Male-female mean test score comparisons have been done for the AFOQT and BAT. The size of the mean differences was expressed in standard deviation units or *d* (Cohen, 1988). The standard deviation for *d* was defined as the within-group standard deviation calculated from the weighted average of the variances for the male and female samples (see for example, McNemar, 1969, p. 115). That is, $SD = (Sp^2/n_1 + Sp^2/n_2)^{1/2}$, where $S_p^2 = (SS_1 + SS_2)/(n_1 + n_2 - 2)$. Thus, $d = (\div_1 - \div_2) / SD$. Cohen (1988) characterizes a *d* of .20 as small, .50 as medium, and .80 as large. However, it should be noted that even "small" *d* values can have a large impact on the proportion of applicants in the lower mean group that would meet or exceed some minimum cut score for selection. Group mean differences were tested using one-tailed t-tests (i.e., males - females) and a .01 Type I error rate. Therefore, positive *d* values indicate higher means for men and negative values indicate higher means for women.

AFOQT. Carretta (1997) examined AFOQT mean score differences in large samples of officer applicants (219,887 men and 50,081 women) and pilot trainees (9,239 men and 237 women). Male officer applicants had

significantly higher means than females on all composites and 15 of 16 tests. The exception was a non-significant effect size of .02 on VA. The mean *d* value for the composites was .442 with a range from .08 (Verbal) to .69 (Pilot). The mean *d* value for the 16 tests was .435 with a range from .02 (VA) to .95 (MC).

Very different results were observed for pilot trainees. The mean Navigator-Technical composite was greater for men than that for women (*d* = .20), but the difference was not nearly as large as for the officer applicants. Means for men did not exceed those for women on the other composites. The mean *d* value for the composites was -.096 and ranged from -.48 (Verbal) to .20 (Navigator-Technical). The mean *d* value for the 16 tests for pilot trainees was .078 and ranged from -.63 (VA) to .84 (MC). The reduction in group differences in the pilot sample was interpreted as being the direct result of the selection process. Means for men exceeded those for women only on the aircrew interest/aptitude tests and some spatial tests: MC (.84), RB (.56), GS (.47), IC (.45), EM (.41), and AI (.22).

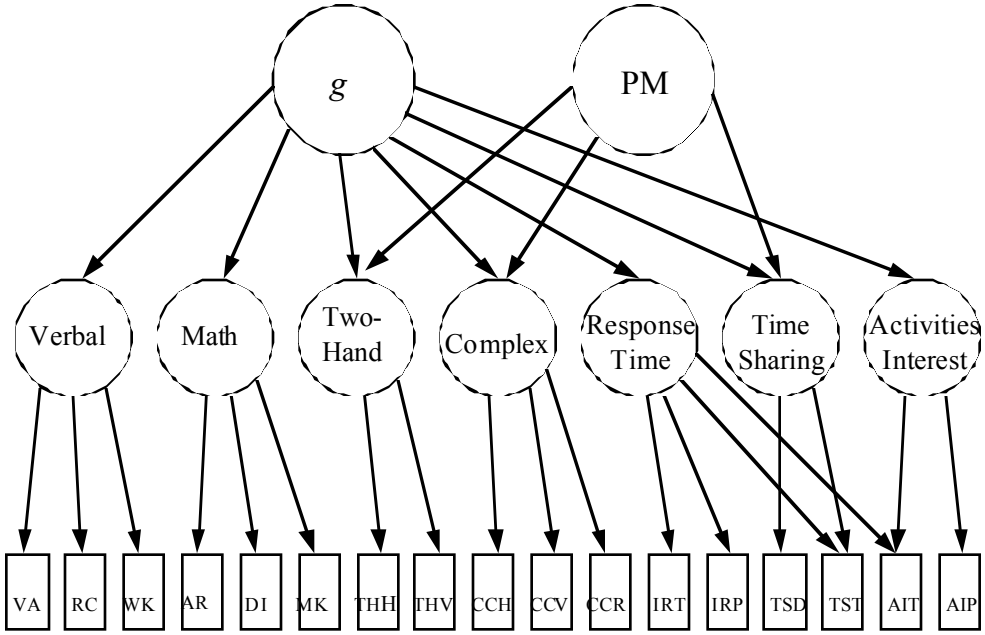

Figure 2. Basic Attributes Test Factor Structure

Note. The higher-order factors are *g* and general psychomotor ability (PM). The lower-order factors are verbal and math from the AFOQT and two-hand coordination, complex coordination, response time, time sharing, and activities interest inventory from the BAT. The AFOQT scores are Verbal Analogies (VA), Reading Comprehension (RC), Word Knowledge (WK), Arithmetic Reasoning (AR), Data Interpretation (DI), and Math Knowledge (MK). The BAT scores are Two-Hand Coordination horizontal and vertical tracking error (THH, THV), Complex Coordination horizontal, vertical, and rudder tracking error (CCH, CCV, CCR), Item Recognition response time and percent correct (IRT, IRP), Time Sharing tracking difficulty and response time (TSD, TSR), and Activities Interest Inventory response time and percent (AIT, AIP).

The results for the officer applicants were consistent with previous findings (Burke, 1995; Hyde, 1981; Jensen, 1980). In a meta-analysis of male-female mean differences on pilot aptitude tests, Burke (1995) reported small differences on verbal tests (-.1 *d* favoring women), with larger differences on quantitative (.5 *d* favoring men) and spatial tests (.5 *d* favoring men). Burke also observed that the size of the sex differences within these broad ability categories varied by specific test content.

BAT. BAT or PCSM scores are not available for officer applicants, as the BAT is taken only by pilot training applicants. A recent analysis of operational BAT data collected between June 1993 and October 1996 revealed small to large mean score differences for 4,888 male and 465 female pilot training applicants. All mean score differences favored men and were statistically significant. The smallest *d* was on Item Recognition (.10), a measure of short-term memory. The largest *d*s were on a psychomotor composite (1.68) that combines scores from Two-Hand Coordination and Complex Coordination and on another psychomotor test called Time Sharing (1.04). The *d* for the PCSM composite which combines the AFOQT Pilot composite, BAT scores, and a flying experience score was .73.

The BAT results are consistent with those reported by Burke (1995) for a sample of U. K. Royal Air Force pilot applicants. He reported a small mean score difference of .02 *d* for 4 information processing tests and a large difference of .98 *d* favoring men for 2 psychomotor tests.

Role of Ability and Prior Job Knowledge on Pilot Skill Acquisition

AFOQT. Ree, Carretta, and Teachout (1995) developed and examined a causal model of the role of *g* and prior flying job knowledge on the acquisition of additional flying knowledge and skills in pilot training. Participants were 3,428 USAF officers attending a 53 week pilot training course. The measures of *g* and prior job knowledge (JK_P) were derived from the AFOQT. Pilot training classroom grades were used to derive the measures of job knowledge acquired during early, middle, and late training (JK_{T1}, JK_{T2}, and JK_{T3}). Pilot training check flight grades were used to produce work sample job performance measures for early and later training (WS_1 and WS_2). The causal model showed that *g* directly influenced the acquisition of job knowledge both prior to and during training. General cognitive ability *indirectly* influenced work sample performance through the acquisition of job knowledge, but did not show any direct influence. Prior job knowledge had almost no influence on subsequent job knowledge, but directly influenced the early work sample. Early training job knowledge influenced subsequent job knowledge and work sample performance. Early work sample performance strongly influenced later work sample performance.

Carretta and Ree (1997) tested the Ree et al. (1995) causal model on separate male (n = 3,369) and female (n = 59) pilot samples. The coefficients for the causal model are shown in Figure 3. The results are considered preliminary due to the small number of women. Results were similar for both sexes. However, the direct and indirect influence of *g* on flying performance was stronger for women than for men. Also, the relationship between prior job knowledge and flying performance was stronger for women. The influence of early flying skills on later flying skills was very strong for both sexes.

BAT. Due to the small number of women tested on the BAT, no studies have been done to examine the causal role of abilities measured by the BAT (e.g., *g* and psychomotor skills) in the acquisition of flying knowledge and skills for men versus women.

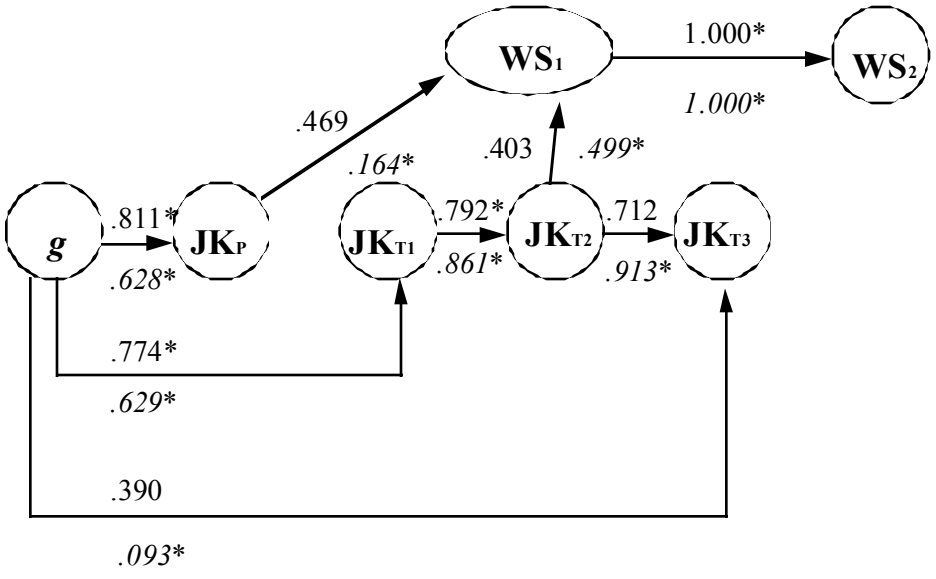

Figure 3. Male-Female Causal Models of Pilot Training

Note. Path coefficients for females are in regular type. Those for males are in italics.

*$p < .01$

Predictive Validity

AFOQT. Roberts and Skinner (1996) investigated the equity of the AFOQT Verbal, Quantitative, and Academic Aptitude composites against Officer Training School (OTS) performance. Participants were 12,166 men and 1,393 women who attended OTS between 1982 and 1988. The criteria were an officer training effectiveness report compiled by course instructors at the 11th week of training and a final course grade based on 5 written tests. Regression analyses indicated level bias with female performance being overpredicted by a small and consistent amount at all aptitude levels.

Three recent studies have investigated the predictiveness of AFOQT scores for male and female pilot trainees (Carretta, 1990, 1997; Siem & Sawin, 1990) and have produced similar results. Carretta (1990) and Siem and Sawin (1990) examined the AFOQT composites, but not the tests. In both studies, higher mean scores were observed for men than for women on pilot selection factors (i.e., Pilot and Navigator-Technical composites) and men were more likely to complete pilot training. However, when men and women were matched on test scores, in most instances, they performed equally well in pilot training. When sex differences occurred, training performance was overpredicted for women.

Carretta (1997) examined the predictive validity of the AFOQT composites and tests for 9,239 male and 237 female pilot trainees. Despite sex differences in AFOQT mean performance, there was no evidence of differential validity. When sex differences in predicted pilot training completion were observed, performance was overestimated for women relative to men. The observed differences in intercepts were eliminated when the regression equations were adjusted for unreliability. No prediction bias was observed against women.

BAT. No studies have been done to examine predictive bias for the BAT and PCSM. The number of pilot trainees that have tested on the operational BAT and completed training is very small. Most pilot applicants test on the BAT after the end of their sophomore year in college. As a result, there usually is about a 2.5 year interval between BAT-testing and completion of pilot training. These data will be examined for predictive bias once enough men and women have completed pilot training.

DISCUSSION

Results from several recent USAF studies of sex differences on officer commissioning and pilot selection tests give a clear picture of gender equity. Studies of the structure of ability have shown very similar results for men and women, despite sex differences in mean score performance. This *factorial invariance* means that the tests are measuring the same thing for both sexes. This is important because it allows us to do other tests of sex differences (e.g., means, validity) without having to worry whether the same constructs are being measured for both sexes.

Mean score differences on USAF pilot selection tests were consistent with previous research (Burke, 1995; Hyde, 1981; Jensen, 1980). Large mean differences favoring men were observed in applicant samples, especially for measures of psychomotor ability, spatial ability, and technical knowledge. Officer and pilot selection procedures reduced, but did not eliminate, sex differences in mean scores. USAF regulations set minimum scores for selection tests and the selection boards use a top-down selection procedure. This has the effect of reducing mean score differences between men and women..

Female applicants were less likely to meet or exceed minimum scores on the AFOQT and BAT. The potential for adverse impact exists to the extent that sex differences on mean test scores occur. It is possible that well qualified women are less inclined to view the Air Force as an attractive career choice. Another possibility is that women are less likely to take courses or pursue leisure interests that might increase their performance on the AFOQT and BAT. Mean differences might be reduced by making information about test content readily available. This is already done for the AFOQT and BAT. Test descriptions and example items are available in free information pamphlets. Those interested in applying for officer commissioning or pilot training can easily determine test content and adopt a suitable preparation strategy. However, some of the tests used for pilot selection rely on flying job knowledge (e.g., AFOQT Aviation Information and Instrument Comprehension tests) that is not readily available or may require a large financial and time investment by the applicant (e.g., completing an aircraft training course).

Despite sex differences in mean test performance, causal models of ability and prior flying knowledge on the acquisition of additional flying knowledge and flying skills showed similar results for men and women. For both sexes, *g* had a direct influence on the acquisition of flying knowledge and an indirect influence on the acquisition of flying skills. The influence of *g* and prior job knowledge on flying performance was stronger for women than for men. The influence of early flying skills on later flying skills was very strong for both sexes.

Results from predictive bias studies of the AFOQT showed no evidence of differential validity. When sex differences in predicted training outcome were observed, performance was overestimated for women relative

to men. The observed differences in intercepts were eliminated when regression equations were adjusted for unreliability. No predictive bias was found. These results were consistent with previous research showing that selection tests that adhere to proper development and administration standards are not more predictive for the majority group than the minority group (Jensen, 1980).

REFERENCES

Bentler, P. M. (1989). *EQS structural equation program manual*. Los Angeles, CA: BMDP Statistical Software.

Berger, F. R., Gupta, W. B., Berger, R. M., & Skinner, J. (1990). *Air Force Officer Qualifying Test (AFOQT) form P: Test manual* (AFHRL-TR-89-56). Brooks Air Force Base, TX: Manpower and Personnel Division, Air Force Human Resources Laboratory.

Burke, E. F. (1995). Male-female differences on aviation selection tests: Their implications for research and practice. In N. Johnston, R. Fuller, & N. McDonald (Eds.), *Aviation Psychology: Training and Selection* (pp. 188-193). Aldershot, England: Avebury Aviation.

Carretta, T. R. (1990, April). *Gender differences in USAF pilot training performance*. Paper presented at the 12th Symposium on Psychology in the Department of Defense, Colorado Springs, CO.

Carretta, T. R. (1992). Recent developments in U. S. Air Force pilot candidate selection and classification. *Aviation, Space, and Environmental Medicine, 63,* 1112-1114.

Carretta, T. R. (1997). Group differences on US Air Force pilot selection tests. *International Journal of Selection and Assessment, 5,* 115-127.

Carretta, T. R., & Ree, M. J. (1993). Basic Attributes Test (BAT): Psychometric equating of a computer-based test. *The International Journal of Aviation Psychology, 3,* 189-201.

Carretta, T. R., & Ree, M. J. (1995). Near identity of factor structure in sex and ethnic groups. *Personality and Individual Differences, 19,* 149-155.

Carretta, T. R., & Ree, M. J. (1996). Factor structure of the Air Force Officer Qualifying Test: Analysis and comparison. *Military Psychology, 8,* 29-42.

Carretta, T. R., & Ree, M. J. (1997). A preliminary evaluation of causal models of male and female acquisition of pilot skills. *The International Journal of Aviation Psychology, 7,* 353-364.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Hyde, J. S. (1981). How large are cognitive gender differences? A meta-analysis using ω^2 and $\underline{d}$. *American Psychologist, 36,* 892-901.

Jensen, A. R. (1980). *Bias in mental testing*. NY: The Free Press.

McArdle, J. J. (1996). Current directions in structural factor analysis. *Current Directions in Psychological Science, 5,* 11-18.

McNemar, Q. (1969). *Psychological statistics* (4th ed.). NY: Wiley.

Ree, M. J., & Carretta, T. R. (1994). The correlation of general cognitive ability and psychomotor tracking tests. *International Journal of Selection and Assessment, 2,* 209-216.

Ree, M. J., & Carretta, T. R. (1995). Group differences in aptitude factor structure. *Educational and Psychological Measurement, 55,* 268-277.

Ree, M. J., Carretta, T. R., & Teachout, M. S. (1995). The role of ability and prior job knowledge in complex

training performance. *Journal of Applied Psychology*, *80*, 721-730.

Roberts, H. E., & Skinner, J. (1996). Gender and racial equity of the Air Force Officer Qualifying Test in officer
training school selection decisions. *Military Psychology*, *8*, 95-113.

Siem, F. M., & Sawin, L. L. (1990, April). *Comparison of male and female USAF pilot candidates.* Paper presented
at the AGARD Symposium on Recruitment, Selection, Training, and Military Operations of Female Aircrew,
Tours, France.